

Twitter 中话题趋势与传播分析

话题趋势分析一直是文本挖掘领域的热点话题，可追溯到话题检测与跟踪（Topic Detection and Tracking, TDT）。TDT 主要是对时序文本流进行分析，与传统的静态文本挖掘不同，TDT 是增量式动态分析，当前的分析结果是基于前面分析结果之上产生的。动态分析的结果可以反映话题或事件的发展趋势，因此在新闻传播领域大量使用。随着微博的出现，话题趋势分析显得更为重要，微博数据量极大，而且存在大量噪声数据，已经无法采取人工方式来分析。然而，微博有着巨大的影响力，有些情况下甚至能超越传统新闻媒体快速传播消息。

在分析微博中话题趋势的同时，还可以分析话题传播的方式以及传播路径。微博群体中通过消息转发和回复进行沟通交流，同时也记录了消息和话题的传播过程。反过来，通过分析话题的传播路径能寻找微博中的群体关系，了解群体关系中哪些角色起到核心作用。我们试图利用文本挖掘技术寻找 Twitter 数据中的一般性的规律，同时将这些规律通过可视化技术呈现，让用户从中发现隐藏的不为人所知的知识。

结合上述两方面，具体问题可以描述如下：

- 1) 分析一段时期内 Twitter 中隐藏哪些话题，趋势如何
- 2) 在话题传播过程中哪些人参与
- 3) 谁在其中起到核心作用
- 4) 参与人对某个话题的观点

从信息技术角度来看，解决上述问题首先需要对问题进行一定程度的抽象。抽象有两种不同类型，一种是任务上的抽象，另一种是数据抽象。任务抽象就是将特定领域的问题转化为一般化的信息处理任务。如“找出 Twitter 中的隐含的话题”其实就是话题检测与跟踪任务；找出传播中的核心人物，就是从传播路径构成的网络中寻找中心度高的节点。数据抽象则是从另外一个角度考虑解决领域问题，将领域中分析的数据抽象为一般化的数据，如高维数据、时变数据，还是层次网络数据等。每一类任务或数据都有一些常规的解决方案，套用这些方案看是否能够解决，如果解决不了，问题在哪里？这样就能产生新的想法和创新。设计解决方案，一方面要充分利用自动处理技术，另外还要结合可视化技术。

回归到上述问题，可以将其概括为方面任务：

- 1) 话题趋势分析
- 2) 与话题相关的社交关系分析
- 3) 与话题相关的舆情分析

■ 话题趋势分析技术

话题是一段时间内人们比较关注的一些问题，如重要事件、基本观念、社会现象等，话题通常具有时间性和地域性。话题趋势分析问题形式化地描述如下：

给定一个时序 post 流 $P_t, t = [1, 2, \dots, \infty)$ ，按时间先后顺序实时到达，确定流中包含的话题以及每个话题关联的 post 的流行程度（即趋势），并监测它们在流行期内的变化。

分析话题趋势基本原则是：1) 寻找回复或转发密集的用户群；2) 对 tweets 分组，不同的组即是不同的话题。前者常用于动态挖掘，后者用于静态分析。分析话题趋势基本路线仍是文本挖掘技术，下面讨论具体的分析技术。

1、跟踪用户感兴趣的关键词

应用统计方法计算单词出现的次数，确定单词在时序上的发展趋势。该方法解决了一个一般性问题：一个关键词或两个关键词有多流行。这种时间趋势分析在一些商业 blog 和 web 搜索引擎中使用较多，如 Google Hot Trends 以及 BlogPlus。但是话题不仅仅是一两个关键词，而是一组有内在关联的关键词。因此这种方法无法真正识别不同的用户关心的话题。

2、基于聚类的话题分析

最近时间特征以及时序演化聚类开始被研究用于社交数据话题趋势监测。例如有人研究不同 blog 之间引用关系构成的图，应用层次聚类算法将语义上关系紧密的 blog 聚集在一起。通过 TF-IDF 方法可以分析每个聚类内部的话题的发展趋势。

3、多种数据源的综合分析

BlogScope 收集来源于 Blogosphere、新闻、社交网络以及其他在线论坛的数据。从时空上监测突发事件。这种方法出发点在于：当一个事件发生时，与该事件相关的 post 会快速增加；事件通常都具有一定的时间和空间范围。TwitterStand 则一方面在线收集 Twitter 中的数据，另外一方面手工选择一些种子数据。手工选择的数据能有效降低 Twitter 数据噪声。

4、基于 burst 模型的趋势分析

TwitterMonitor 与 Blogscope 类似在线监测 Twitter 中的趋势，整个过程分为两步，首先监测一些突发的关键词，根据它们出现的频率；然后对每个爆发的关键词获取相关的历史 tweets，关键词根据它们共同出现的次数进行聚类组成趋势。

5、面向多关键词的事件监测

通常只分析单个词的出现频率，根据出现频率计算 burst 数量，对 term 进行聚类来监测事件。本文的出发点略有不同，分析多词，因为多个词语义上更能反映真实事件，如 Argentina vs Nigeria 就要比两个单独的词语义更明确，可能是两个国家之间的比赛。因此本文采用了基于 segment 的方法监测 Twitter 中的事件。一个 segment 是指连续出现的一个或多个单词，由 segment 组成的事件通常具有自解释性，如由 5 个 segment 组成的事件[south korea, greece, korea vs greece, korea won, korea] on 12 June 2010。整个分析过程分为 5 个阶段：

- 1) 调用 Microsoft 的 N-Gram 服务对 tweets 进行 segment；
- 2) 识别突发的 segment；
- 3) 进一步通过用户对这些 segment 的关注程度（user frequency），确认是否为事件相关的 segment（Event Segment）；
- 4) 对 Event Segments 进行聚类，分组后的 segment 成为候选事件（Candidate Event）；
- 5) 通过计算 Candidate Event 中 segment 相似度确定 Event；
- 6) 通过与 Wikipedia 中真实事件对比，确定分析结果是否正确。

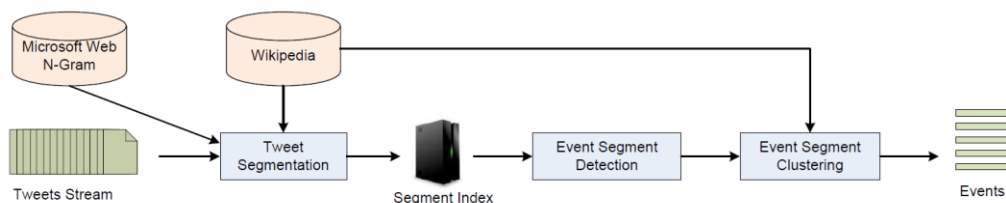


Figure 1: Segment-based Event Detection System Architecture

一般情况下，话题趋势分析都包含两个重要阶段：1) 监测话题 2) 话题分类。监测话题首先要监控是否存在一些频率比较高的词，但是某些高频率的词可能是一些无意义的词，如类似 http, cnn 等。因此还需要验证这些词是否是话题中的词。要么是与其他有明确含义的信息源对比，要么是手工选择一些 tweets，要么是看这些词是否真正被用户经常关注。分别对应上述三种方法（3、4、5）。

但是这些方法存在一些问题，简单的方法效果不好，而复杂的方法效果较好但是实现困难。今天我联系了方法 5 的作者（南洋理工大学博士），他同意给我一些帮助。

下一步，在确定话题及趋势分析之后，就是要分析话题中参与人之间的社交关系。这里的方法相对成熟，而且不需要太复杂，可以直接可视化。

如果上述两方面都能走通，再结合一些较好的数据，如‘斯诺登’事件，就可以分析一些有趣的结果了。

下周，寻找比较简单有效的话题分析方法，并抓取最近关于‘斯诺登’事件的 **Tweets**。其实本周前段时间都是在做 **Twitter** 数据分析，分别尝试了直接的 **TF-IDF** 方法、以及命名实体提取方法，虽然能看到一些零散的词，但是总体效果并不理想，所以找来最近的文献看，才发现微博分析还是一个较难的问题，最近研究很热，上述分析技术都是最近两年的论文。

参考文献

- [1] Becker, H., Mor Naaman, and Luis Gravano (2011). Beyond Trending Topics: Real-World Event Identification on Twitter. ICWSM.
- [2] Vakali, A., et al. (2012). Social networking trends and dynamics detection via a cloud-based framework design. Proceedings of the 21st international conference companion on World Wide Web. Lyon, France, ACM: 1213-1220.
- [3] Li, C., et al. (2012). Twevent: segment-based event detection from tweets. Proceedings of the 21st ACM international conference on Information and knowledge management. Maui, Hawaii, USA, ACM: 155-164.
- [4] Mathioudakis, M. and N. Koudas (2010). TwitterMonitor: trend detection over the twitter stream. Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. Indianapolis, Indiana, USA, ACM: 1155-1158.
- [5] Alvanaki F, Sebastian M, Ramamritham K, et al. EnBlogue: emergent topic detection in web 2.0 streams[C]//Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, 2011: 1271-1274.
- [6] Shan D, Zhao W X, Chen R, et al. Eventsearch: A system for event discovery and retrieval on multi-type historical data[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 1564-1567.